

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号  
特開2003-271592  
(P2003-271592A)

(43) 公開日 平成15年9月26日 (2003.9.26)

(51) Int.Cl. <sup>7</sup>	識別記号	F I	テームト* (参考)
G 0 6 F 17/28		C 0 6 F 17/28	R 5 B 0 7 j
17/30	1 7 0	17/30	1 7 0 A 5 B 0 9 1
	3 2 0		3 2 0 D

審査請求 有 請求項の数10 O L (全 9 頁)

(21) 出願番号 特願2002-74270 (P2002-74270)

(22) 出願日 平成14年3月18日 (2002.3.18)

(71) 出願人 301022471

独立行政法人通信総合研究所  
東京都小金井市貫井北町4-2-1

(72) 発明者 内元 清貴

東京都小金井市貫井北町4-2-1 独立  
行政法人通信総合研究所内

(74) 代理人 100090893

弁理士 渡邊 敏

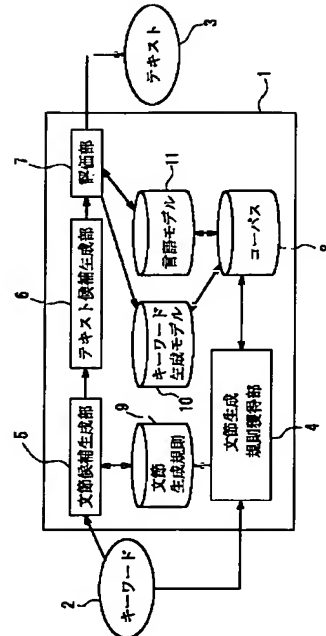
最終頁に続く

(54) 【発明の名称】 テキスト生成方法及びテキスト生成装置

(57) 【要約】

【課題】 入力するキーワードが十分でない場合にも、有意なテキストを生成するテキストの生成方法・生成装置を提供すること。

【解決手段】 キーワード2をテキスト生成装置1に入力し、文節生成規則獲得部4ではコーパス8から文節生成規則9を獲得し、文節候補生成部5でキーワード2から文節候補を生成する。文節候補に係り受け関係を仮定してテキスト候補生成部6でテキスト候補を生成し、評価部7で評価付けし、評価に従ってテキスト3を出力する。



【特許請求の範囲】

【請求項1】所定の言語の文又は文章のテキストを生成するテキスト生成方法であって、

1個以上のキーワードとなる単語を入力する入力ステップ、  
キーワードから文字単位候補を生成する文字単位候補生成ステップ、  
該各文字単位候補の係り受け関係を仮定してテキスト候補を生成するテキスト候補生成ステップ、  
該テキスト候補を評価付けする評価ステップ、  
評価付けされた少なくとも1つのテキスト候補を出力する出力ステップの各ステップを含むことを特徴とするテキスト生成方法。

【請求項2】前記入力ステップにおいて、  
当該言語のデータベースを用いて入力された単語と係り受け関係を有する単語を新たなキーワードとして入力する請求項1に記載のテキスト生成方法。

【請求項3】前記文字単位候補生成ステップにおいて、  
少なくとも1つの該キーワードの前又は後に、該キーワードに関連した文字列を付加して文字単位候補を生成し、他の全ての該キーワードについても同様に文字列を付加し、或いは付加せずに文字単位候補を生成する請求項1又は2に記載のテキスト生成方法。

【請求項4】前記テキスト生成方法において、入力ステップの次に、  
該キーワードを含む文・語句を、データベースから抽出する抽出ステップ、  
抽出された文・語句から文字単位候補の生成規則を自動獲得する生成規則獲得ステップを備え、  
文字単位候補生成ステップにおいて、該生成規則を用いて文字単位候補を生成する請求項1ないし3に記載のテキスト生成方法。

【請求項5】前記生成規則獲得ステップが、  
前記抽出ステップで抽出された文・語句を形態素解析及び／又は構文解析を行い、  
解析された前記キーワードを含む文字単位を生成規則とする請求項4に記載のテキスト生成方法。

【請求項6】所定の言語の文又は文章のテキストを生成するテキスト生成装置であって、  
1個以上のキーワードとなる単語を入力する入力手段と、  
キーワードから文字単位候補を生成する文字単位候補生成手段と、  
該各文字単位候補の係り受け関係を仮定してテキスト候補を生成するテキスト候補生成手段と、  
該テキスト候補を評価付けする評価手段と、  
評価付けされた少なくとも1つのテキスト候補を出力する出力手段とを少なくとも備えることを特徴とするテキスト生成装置。

【請求項7】前記入力手段において、

当該言語のデータベースを用いて入力された単語と係り受け関係を有する単語を新たなキーワードとして入力する請求項6に記載のテキスト生成装置。

【請求項8】前記文字単位候補生成手段において、  
少なくとも1つの該キーワードの前又は後に、該キーワードに関連した文字列を付加して文字単位候補を生成し、他の全ての該キーワードについても同様に文字列を付加し、或いは付加せずに文字単位候補を生成する請求項6又は7に記載のテキスト生成装置。

【請求項9】前記テキスト生成装置であって、  
入力手段で入力されたキーワードを含む文・語句を、データベースから抽出する抽出手段と抽出された文・語句から文字単位候補の生成規則を自動獲得する生成規則獲得手段とを備え、  
文字単位候補生成手段において、該生成規則を用いて文字単位候補を生成する請求項6ないし8に記載のテキスト生成装置。

【請求項10】前記生成規則獲得手段が、  
前記抽出手段で抽出された文・語句を形態素解析及び／又は構文解析を行い、  
解析された前記キーワードを含む文字単位を生成規則とする請求項9に記載のテキスト生成装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は自然言語処理の方法及び装置に関する。特に、いくつかのキーワードからテキストを生成する手法に関わる。

【0002】

【従来の技術】テキスト生成は機械翻訳、要約、対話システムなど自然言語処理の様々な応用に利用される重要な要素技術の一つである。近年、大量のコーパスが利用可能となり、自然な表層文を生成する目的にもコーパスが利用されるようになってきた。その典型例の一つが原言語から目的言語への機械翻訳に用いられる言語モデルである。

【0003】例えば、本件出願人が特願2001-395618号で開示したテキスト生成のシステムでは、置き換えた単語や句を目的言語側で尤もらしい順序に並び替え、目的言語を生成する。言語モデルの入力は、一般に語の集合であり、言語モデルに要求されるのは、基本的にそれらの語の並べ換えである。このような従来のシステムでは、与えられた語の集合を並べ換えると自然な文を生成できるという仮定がある。つまり、自然な文を生成するための語の集合は翻訳モデルにより過不足なく生成されることが前提となっている。

【0004】しかし、この前提のためには大規模な対訳コーパスが必要であり、日本語などの比較的コーパスが整備された言語が原言語であっても、対象言語との対訳コーパスの状況、対象言語におけるコーパスの状況によっては、上記従来の手法では十分なテキスト生成が行え

ない場合があった。また、上記開示でもある程度の語句の補完は行うが、補助的な補完を行うのみで、効率的に関連する語句を補完することはできなかった。

【0005】この問題は機械翻訳に限らず、一般的にテキスト生成において生じる問題であり、原言語テキストが完全なものでなく、誤りのあるOCR 認識結果や音声認識結果などの場合には同様に高精度なテキスト生成ができない問題があった。

【0006】

【発明が解決しようとする課題】本発明は、上記従来技術の有する問題点を鑑みて創出されたものであり、その目的は、入力するキーワードが十分でない場合にも、有意なテキストを生成するテキストの生成方法・生成装置を提供することである。

【0007】

【課題を解決するための手段】本発明は、上記の課題を解決するために、次のようなテキスト生成方法を創出する。すなわち、日本語など所定の言語の文又は文章のテキストを生成するテキスト生成方法において、まず入力ステップで1個以上のキーワードとなる単語を入力する。そして、キーワードから文字単位候補を生成する文字単位候補生成ステップ、該各文字単位候補の係り受け関係を仮定してテキスト候補を生成するテキスト候補生成ステップ、該テキスト候補を評価付けする評価ステップ、評価付けされた少なくとも1つのテキスト候補を出力する出力ステップの各ステップを含む。

【0008】前記入力ステップにおいて、当該言語のデータベースを用いて入力された単語と係り受け関係を有する単語を新たなキーワードとして入力することで、キーワードが十分でない場合に補完させることもできる。

【0009】文字単位候補生成ステップにおいて、少なくとも1つの該キーワードの前又は後に、該キーワードに関連した文字列を付加して文字単位候補を生成してもよい。その際、他の全ての該キーワードについては、同様に文字列を付加してもよいし、或いは付加しないで文字単位候補を生成してもよい。

【0010】また、本発明によるテキスト生成方法において、入力ステップの次に、キーワードを含む文・語句を、データベースから抽出する抽出ステップ、抽出された文・語句から文字単位候補の生成規則を自動獲得する生成規則獲得ステップを備え、文字単位候補生成ステップにおいて、該生成規則を用いて文字単位候補を生成する構成でもよい。ここで、データベースには例えば日本語など、本発明で生成するテキストの言語のコーパスを用いることができる。

【0011】さらに、生成規則獲得ステップが、抽出ステップで抽出された文・語句を形態素解析及び／又は構文解析を行い、解析された前記キーワードを含む文字単位を生成規則とする構成でもよい。

【0012】さらに本発明は、所定の言語の文又は文章

のテキストを生成するテキスト生成装置を提供することでもできる。該装置には、1個以上のキーワードとなる単語を入力する入力手段と、キーワードから文字単位候補を生成する文字単位候補生成手段と、各文字単位候補の係り受け関係を仮定してテキスト候補を生成するテキスト候補生成手段と、テキスト候補を評価付けする評価手段と、評価付けされた少なくとも1つのテキスト候補を出力する出力手段とを少なくとも備える。

【0013】ここで、入力手段において、当該言語のデータベースを用いて入力された単語と係り受け関係を有する単語を新たなキーワードとして入力してもよい。また、文字単位候補生成手段において、少なくとも1つの該キーワードの前又は後に、該キーワードに関連した文字列を付加して文字単位候補を生成し、他の全ての該キーワードについても同様に文字列を付加し、或いは付加せずに文字単位候補を生成する構成をとってもよい。

【0014】本発明によるテキスト生成装置にさらに、入力手段で入力されたキーワードを含む文・語句を、データベースから抽出する抽出手段と、抽出された文・語句から文字単位候補の生成規則を自動獲得する生成規則獲得手段とを備えてもよく、文字単位候補生成手段において、該生成規則を用いて文字単位候補を生成する構成をとることもできる。

【0015】生成規則獲得手段が、抽出手段で抽出された文・語句を形態素解析及び／又は構文解析を行い、解析された前記キーワードを含む文字単位を生成規則としてもよい。

【0016】

【発明の実施の形態】以下、本発明の実施方法を図面に示した実施例に基づいて説明する。なお、本発明の実施形態は以下に限定されず、適宜変更可能である。図1には本発明におけるテキスト生成装置（以下、本装置）

(1)の説明図を示す。最も単純な本装置(1)の機能として、例えば「彼女」「家」「行く」の3つのキーワード(2)が入力された時に、「彼女の家に行く」(3a)「彼女が家に行った」(3b)などのテキストを生成する。

【0017】本装置(1)の具体的な構成例として図2に示す各部を備える。本装置(1)は例えば、CPUとメモリ、ハードディスクなどの外部記憶媒体を備えるパーソナルコンピュータなどにより構成することができ、主な処理をCPUにおいて行い、処理の結果を随時RAM、外部記憶媒体に記録する。本発明において入力となるキーワード(2)は、パーソナルコンピュータに接続されたキーボードを用いて入力したり、他の言語処理システムから出力されたデータを用いることができる。

【0018】本実施例で、キーワード(2)は2つの処理に用いられる。その1つは文節生成規則獲得部(4)であり、もう1つは文節候補生成部(5)である。ここでは、日本語を対象とし、生成する文字単位は文節であ

る。キーワードは文節の主辞となる語であると定義する。そして、文節の主辞となる語は、文末に一番近い内容語であるとする。ここで、内容語は、その語の品詞が、動詞、形容詞、名詞、指示詞、副詞、接続詞、連体詞、感動詞、未定義語である形態素の見出し語であるとし、それ以外の形態素の見出し語を機能語とする。ただし、サ変動詞、動詞「なる」、形式名詞「の」については、文節内で他に内容語がない場合を除いて機能語として扱う。品詞の体系は京大コーパス(Version3.0) (黒橋長尾1997)のものに従った。

【0019】文節生成規則獲得部(4)では、キーワード「彼女」、「家」、「行く」が与えられたとき、それぞれを含む文をコーパス(8)から検索し、形態素解析、構文解析(係り受け解析)をする。そして、そこからキーワード(2)を含む文節を抽出して、キーワードから文節を生成する規則「彼女」→「彼女の」、「彼女」→「彼女が」、「家」→「家に」、「行く」→「行く」、「行く」→「行った」などの文節生成規則(9)を獲得し、記録する。

【0020】ここで、生成規則の自動獲得には次の手法を用いる。キーワードの集合を $V$ とし、キーワード $k(k \in V)$ から文節を生成する規則の集合を $R_k$ とすると、規則 $r_k(k \in R_k)$ は次の形式で表現されるものと定義する。

$$k \rightarrow h_k m^*$$

ここで、 $h_k$ はキーワードを含む主辞形態素、 $m^*$ は同じ文節内で $h_k$ に連続する任意個の形態素とする。キーワードが与えられると、この形式を満たす規則を単言語コーパスから自動獲得する。

【0021】一方、文節候補生成部(5)では、文節生成規則(9)を参照しながら、入力されたキーワード(2)から出力するテキスト(3)を構成する文節の候補を生成する。例えば、「彼女」では自然なテキストを構成する文節とはなりにくいが、「彼女の」あるいは「彼女が」のように「彼女」という単語と極めて密接な関連性を有する語句を付加し、後段の処理によるテキスト生成に備える。

【0022】本実施例のように、文節生成規則獲得部(4)によりコーパス(8)から入力するキーワード(2)の文節規則を生成することで、最小限の計算量で効果的に文節生成規則を得ることができ、処理速度の向上に寄与する。

【0023】しかしながら、本発明の実施においては必ずしもキーワード(2)に関連する語句をコーパスから抽出する構成を取る必然性はなく、計算能力に応じて任意の語句を入力されたキーワード(2)の前後に付加してもよい。本発明では後述の評価部(7)により、任意の語句を付加しても当該文節候補について緻密な評価がされるため、これにより最も評価値の高くなる文節候補が生成できるようになる。

【0024】次に、テキスト候補生成部(6)でテキス

ト候補を生成する。テキスト候補はグラフあるいは木の形で表現する。すなわち、図3に示すように、各文節候補(4aないし4f)の間に係り受けの関係を仮定して、テキスト候補1(12)、テキスト候補2(13)のような文節を単位とした依存構造木の形でテキスト候補を生成する。

【0025】このとき、次の条件を満たすように依存構造木の候補を生成する。

- (i) 係り受けは前方から後方に向いている。(後方修飾)
- (ii) 係り受け関係は交差しない。(非交差条件)
- (iii) 係り要素は受け要素を一つだけ持つ。

例えば、キーワードが3個の場合、キーワードを含む文節候補がそれぞれ $b_1$ 、 $b_2$ 、 $b_3$ であったとすると、順序を固定した場合には、 $(b_1(b_2 b_3))$ 、 $((b_1 b_2) b_3)$ の2通り、固定しない場合には16通りの候補ができる。

【0026】生成されたテキスト候補(12・13など)は、評価部(7)でコーパスから学習したキーワード生成モデル(10)や言語モデル(11)を用いて順序付けされる。以下、キーワード生成モデル(10)と、言語モデル(11)として形態素モデル及び係り受けモデルについて説述する。

【0027】キーワード生成モデルでは、次の5種類の情報を素性として用いたモデル(KM1ないし5)を考える。以下で、キーワードの集合 $V$ は、ある回数以上コーパスに出現した主辞単語の集合とし、文節は前記で表現されるものと仮定する。また、各キーワードは独立であり、与えられたテキストが単語列 $w_1 \cdots w_m$ からなるとき、キーワード $k_i$ は単語 $w_j(1 \leq j \leq m)$ に対応していると仮定する。

【0028】

【KM1】前方の二単語を考慮(trigram)

$k_i$ は前方の二単語 $w_{j-1}$ と $w_{j-2}$ のみに依存すると仮定する。

【式1】

$$P(K|M, D, T) = \prod_{i=1}^n P(k_i | w_{j-1}, w_{j-2})$$

【0029】

【KM2】後方の二単語を考慮(後方trigram)

$k_i$ は後方の二単語 $w_{j+1}$ と $w_{j+2}$ のみに依存すると仮定する。

【式2】

$$P(K|M, D, T) = \prod_{i=1}^n P(k_i | w_{j+1}, w_{j+2})$$

【0030】

【KM3】係り文節を考慮(係り文節)

$k_i$ を含む文節に係る文節がある場合、 $k_i$ はそのうち最も文末側の文節の末尾から二単語 $w_l$ と $w_{l-1}$ のみに依存すると仮定する(図4参照)。

【式3】

$$P(K|M, D, T) = \prod_{i=1}^n P(k_i|w_i, w_{i-1})$$

【0031】

【KM4】受け文節を考慮(受け文節)

$k_i$  を含む文節を受ける文節がある場合、 $k_i$  はその文節内の主辞単語から二単語 $w_s$  と $w_{s+1}$  のみに依存すると仮定する(図4参照)。

【式4】

$$P(K|M, D, T) = \prod_{i=1}^n P(k_i|w_s, w_{s+1})$$

【0032】

【KM5】係り文節を最大二文節考慮(係り二文節)

$k_i$  を含む文節に係る文節がある場合、 $k_i$  は、そのうち最も文末側の文節の末尾から二単語 $w_l$ 、 $w_{l-1}$  と、最も文頭側の文節の末尾から二単語 $w_h$ 、 $w_{h-1}$  のみに依存すると仮定する(図4参照)。

【式5】

$$P(K|M, D, T) = \prod_{i=1}^n P(k_i|w_l, w_{l-1}, w_h, w_{h-1})$$

【0033】次に、形態素モデル(MM)について示す。形態素に付与するべき文法的属性が1個あると仮定する。テキストつまり文字列が与えられたとき、その文字列が形態素であり、かつ $j$  ( $1 \leq j \leq l$ ) 番目の文法的属性を持つとしたときの尤もらしさを確率値として求めるモデルを用いる。テキスト $T$  が与えられたとき、順序付き形態素集合 $M$  が得られる確率は、各形態素 $m_i$  ( $1 \leq i \leq n$ ) が独立であると仮定し、

【式6】

$$P(M|T) = \prod_{i=1}^n P(m_i|m_1^{i-1}, T)$$

と表す。ここで、 $m_i$  は1 から $l$  までのいずれかの文法的属性を表わす。

【0034】一方、係り受けモデル(DM)は、テキスト $T$ と順序付き形態素集合 $M$ が与えられたとき、各文節に対する係り受けの順序付き集合 $D$ が得られる確率は、各々の係り受け $d_1 \cdots d_n$  が独立であると仮定し、

【式7】

$$P(D|M, T) = \prod_{i=1}^n P(d_i|M, T)$$

と表わす。

【0035】例えば、「彼女 公園 行った」の3つのキーワードから「(彼女は(公園へ行った))」と「((彼女の公園へ) 行った)」の2つの候補が生成されたとする。係り受けモデルにより、このうち尤もらしい係り受

け構造を持つ候補が優先される。

【0036】以上に示すような各モデルを用い、本発明では評価部(7)においてテキスト候補(12・13など)に評価付けを行う。そして、評価値が最大あるいは閾値を超えるテキスト候補、あるいは評価値の上位 $N$ 個を表層文に変換して出力する。

【0037】出力方法としては、モニタによる表示の他、音声合成を用いた発声、翻訳システムなど他の言語処理システムへのデータ出力などが可能である。これにより、例えばキーワード(2)として「彼女」「家」「行く」を入力したときに、「彼女の家に行く」(3a)「彼女が家に行った」(3b)などのテキスト(3)を出力することができる。なお、出力は上記した通り、最も評価値の高いものを1つ選択してもよいし、評価値順に複数出力してもよく、例えば複数提示してキーワードを入力した者が最適なものを選択するようにしてもよい。

【0038】以上に示した実施例では、キーワードの前後に語句を付加する構成を主としているが、本発明の実施においてはキーワード(主辞単語に相当するもの)そのものを補完する構成をとることもできる。例えば、「彼 本」から述語などを補完して「彼が本を読んだ」や「彼が本を書いた」、「彼が本を買った」などを生成するために、入力されたキーワードに関連する新たなキーワードを追加して入力することも可能である。

【0039】具体的な実施例としては、図2の構成に図5の要素を追加する。すなわち、キーワード(2)に係り受け関係語抽出部(14)にも入力し、該部(14)ではコーパス(8)から該キーワード(2)と係り受け関係にある単語を抽出する。そして、単語を新たなキーワードとして加え、もともと入力されたキーワード(2)と合わせて文節候補生成部(5)における処理を行う。

【0040】例えば、「(彼が(本を 読む))」そのものがコーパス(8)に無くとも、「(彼が 読む)」と「(本を 読む)」という係り受け関係がそれぞれコーパス(8)にあれば、それらに共通する単語「読む」を新たにキーワードとして追加することによって、文節候補生成部(5)によって「読んだ」が生成できるようになる。

【0041】本構成は、計算量が少なく高速なキーワードの追加が可能であるが、本発明では必ずしもコーパスから係り受け関係にある単語を抽出することに限らず、任意のキーワードの候補を追加し、その中から評価部(7)における評価が結果的に最も高くなるものを出力してもよい。これによって、キーワードにテキストの意味を決定する重要な単語が一部欠落していたとしても、有意なテキストが出力できるようになる。

【0042】なお、上記実施例では生成するテキストを日本語とし、キーワードから生成する文字単位を文節と

していたが、本発明は任意の言語に適用可能である。例えば、英語の場合、複数の名詞句、動詞句が集まって別の名詞句や動詞句を形成し、階層をなす場合がある。そのような言語では、中でも最も小さな句、「基本句」を

文節の代わりに用いることもできる。

【0043】

【表1】

入力(キーワード)			システムの出力例
ロシア	連覇	達成	((ロシアで(連覇を)達成した))
世界	記録	変わる	((世界の(記録に)変わる))
課題	実現	挑む	((課題の(実現に)挑む))
チーム	ともに	勝つ	((チームで(ともに)勝った))
この	計画	決める	
戦後	生活	支える	((戦後の(生活を)支える))
大統領	会見	陳述	((大統領の(会見の)陳述))
容疑者	中国	逮捕	((容疑者は(中国で)逮捕された))
日本	五輪	開催	
国	政策	発足	((国の(政策が)発足する))
課題	実現	向かう	((課題の(実現に)向かう))
社会	働く	女性	((社会に(働いている)女性))
首相	政権	奪回	((首相が(政権を)奪回する))
提案	見直し	作る	((提案の(見直しを)作る))
政権	発足	決まる	((政権が(発足に)決まった))
香港	クルー	多い	((香港は(クルーも)多いらしい))
大会	出場	勝つ	
正月	旅行	急増	((正月の(旅行も)急増している))
民営化	反対	人	((民営化に(反対している)人))
昨年	政策	認める	((昨年の(政策を)認めるべきだ))
米国	勝つ	明らかだ	((米国に(勝てば)明らかになる))
住民	疑い	広がる	((住民の(疑いが)広がる))
日本	中国	近い	
外国人	加入	増加	((外国人の(加入者が)増加している))
五輪	選手権	獲得	((五輪の(選手権を)獲得する))
企業	撤退	進める	((企業が(撤退を)進める))
将来	新造党	生まれる	((将来は(新造党が)生まれるだろう))
昨年	記録	上回る	((昨年の(記録を)上回る))
強い	チーム	目指す	
よい	仕事	ほしい	

【0044】最後に、本発明によるテキスト生成方法を用いた実験例を示す。実験は表1に示す3つの入力するキーワードで行い、出力されたテキストが主観的に正しいか否かで評価した。評価基準は以下の2つである。

【0045】基準1： 1位の候補が意味的、文法的に適切であればシステムの出力が正しいと判断する。

基準2： 上位10位に意味的、文法的に適切な候補があればシステムの出力が正しいと判断する。

評価結果を表2に示す。

【0046】

【表2】

モデル	基準 1	基準 2
KM1 (trigram)	13/30	28/30
KM1 + MM	21/30	28/30
KM1 + DM	12/30	28/30
KM1 + MM + DM	26/30	28/30
KM2 (後方 trigram)	6/30	15/30
KM2 + MM	8/30	20/30
KM2 + DM	10/30	20/30
KM2 + MM + DM	9/30	25/30
KM3 (係り文節)	13/30	29/30
KM3 + MM	26/30	29/30
KM3 + DM	14/30	28/30
KM3 + MM + DM	27/30	29/30
KM4 (受け文節)	10/30	18/30
KM4 + MM	9/30	26/30
KM4 + DM	9/30	22/30
KM4 + MM + DM	13/30	27/30
KM5 (係り二文節)	12/30	26/30
KM5 + MM	17/30	28/30
KM5 + DM	12/30	27/30
KM5 + MM + DM	26/30	28/30

【0047】生成規則により生成されたテキスト候補の数は、2つのキーワードが入力の場合、キーワード一組あたり平均868.8個(26,064/30)、3つのキーワードが入力の場合、キーワード一組あたり平均41,413.5個(1,242,404/30)であった。表2では、前述したキーワード生成モデル(KM1ないし5)と言語モデル(MM及びDM)について、+は各モデルの組み合わせを表わしている。

【0048】表2から分かるように、KM1やKM3、KM5のモデルにMMとDMを組み合わせた場合が最も良い結果となった。MMやDMを用いた場合、用いなかった場合と比べて基準1による評価結果が飛躍的に良くなっているが、その理由は、名詞と格の結び付きより、動詞と格の結び付きの方が強く、後者に着目して学習しているKM1やKM3、KM5のモデルが潜在的に自然な文となる候補を上位に順序付けていたからである可能性が高いと考えられる。

【0049】以上の結果から、本発明の評価部(7)では上述したキーワード生成モデル1, 3, 5と、形態素モデル、係り受けモデルを組み合わせて評価部を構成するのが望ましく、とりわけKM3を用いると特に好適である。これらの組み合わせでは、表2に見るように9割前後の割合で正しいテキストを生成することに成功した。

【0050】

【発明の効果】本発明は、以上の構成を備えるので、次の効果を奏する。すなわち、請求項1ないし5に記載のテキスト生成方法によれば、従来のテキスト生成方法では困難であった入力するキーワードが十分でない場合にも、有意なテキストを生成することができる。

【0051】特に、請求項2に記載のテキスト生成方法では、キーワードと係り受け関係を持つ単語を抽出し、キーワードを追加することができるため、より広義のテキスト生成が実現可能となる。

【0052】あるいは、請求項4に記載のテキスト生成方法では、抽出された文・語句から文字単位候補の生成規則を自動獲得することができるため、効率的に文字単位候補を生成することができ、処理の高速化に寄与する。

【0053】さらに本発明では請求項6ないし10に記載のテキスト生成装置によって、従来のテキスト生成方法では困難であった入力するキーワードが十分でない場合にも、有意なテキストを生成可能である。

【0054】特に、請求項7に記載の発明では、キーワードと係り受け関係を持つ単語を抽出し、キーワードを追加することができるため、より広義のテキスト生成を実現するテキスト生成装置が提供できる。

【0055】請求項9に記載のテキスト生成装置では、抽出された文・語句から文字単位候補の生成規則を自動獲得することができるため、効率的に文字単位候補を生成することができ、処理の高速化、低コスト化に寄与する。本発明は、上記のように優れたテキスト生成方法を提供するテキスト生成装置を創出し、自然言語処理技術の向上に寄与することが出来る。

【図面の簡単な説明】

【図1】本発明によるテキスト生成装置の説明図である。

【図2】同、構成図である。

【図3】キーワードからのテキスト生成の例を示す説明図である。

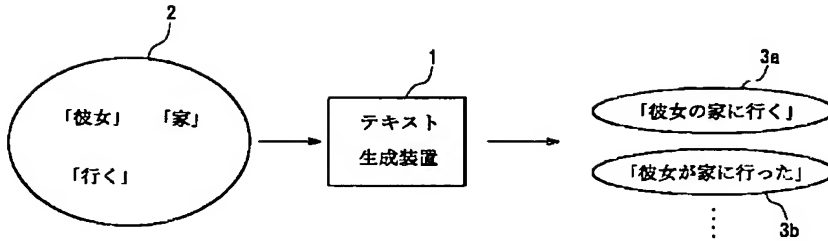
【図4】キーワードと単語の関係を示す説明図である。

【図5】本発明に係る係り受け関係語抽出部の構成図である。

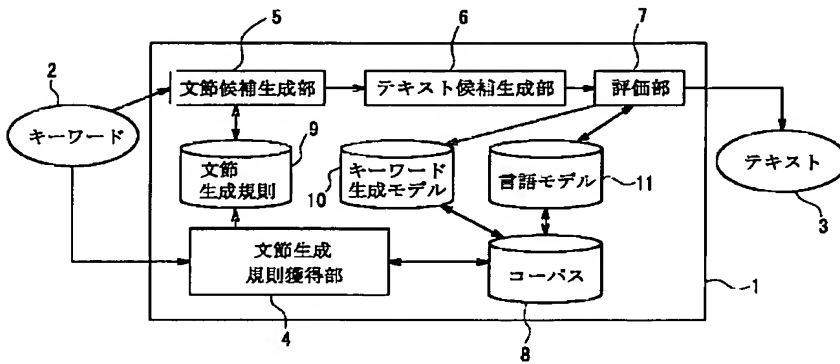
【符号の説明】

- 1 テキスト生成装置
- 2 キーワード
- 3 テキスト
- 4 文節生成規則獲得部
- 5 文節候補生成部
- 6 テキスト候補生成部
- 7 評価部
- 8 コーパス
- 9 文節生成規則
- 10 キーワード生成モデル
- 11 言語モデル

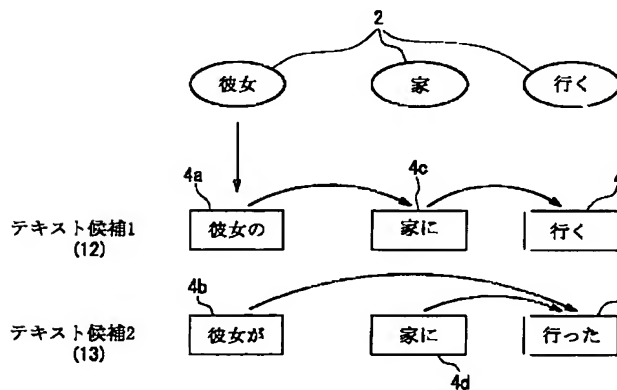
【図1】



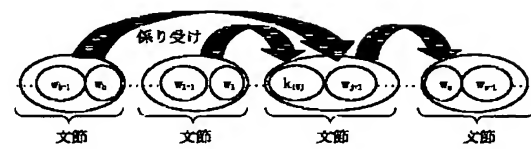
【図2】



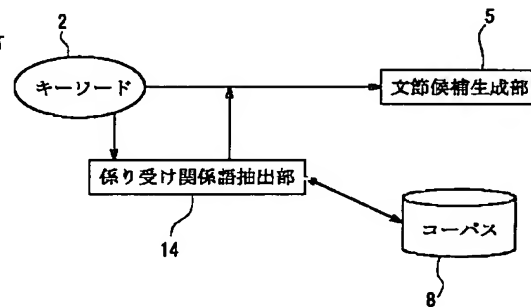
【図3】



【図4】



【図5】





フロントページの続き

(72)発明者 関根 聡

アメリカ合衆国、ニューヨーク州10003  
ニューヨーク、セブンスフロアー、ブロー  
ドウェイ、715、ニューヨークユニバーシ  
ティ、コンピュータサイエンスデパートメ  
ント

(72)発明者 井佐原 均

東京都小金井市貫井北町4-2-1 独立  
行政法人通信総合研究所内

Fターム(参考) 5B075 ND03 NK02 NK32 NK35 PP02  
PP22  
5B091 AA15 CA21 CA24 CC01 CC03  
CC15